评估零效应的三种统计方法

许岳培[1,2*]，陆春雷[3*], 王珺[4*]，宋琼雅[4*]，贾彬彬[5*]，胡传鹏[6]

（[1] 中国科学院行为科学重点实验室（中国科学院心理研究所）, 北京100101；

[2] 中国科学院大学心理学系, 北京100049；

[3] 浙江师范大学心理与脑科学研究院, 金华 321004；

[4] 中山大学心理学系，广州，510006；

[5] 上海体育学院，上海，200438；

[6] 南京师范大学心理学院, 南京, 210097）

注：标星号(*)的作者对本文有同等贡献。作者贡献：王珺、贾彬彬负责贝叶斯估计部分的原理与实例分析；宋琼雅、许岳培负责等价性检验部分的原理与实例分析；陆春雷负责前言、最小感兴趣效应量区间、贝叶斯因子实例分析和三种方法的比较；另外，许岳培还负责模拟比较，贾彬彬还负责数据可视化；胡传鹏提出想法并指导论文的写作。所有人均阅读并同意本文的最终版本。感谢中国科学院心理研究所的李云箫同学对文本的建议和修改。

通讯作者：胡传鹏, Email: hcp4715@hotmail.com

评估零效应的三种统计方法

摘 要 在心理学研究中，以下两种情况下研究者需要评估效应是否*不存在*:（1）研究设计或者假设中需要证明某个效应不存在;（2）研究者本意是要拒绝零效应但未能拒绝（即意外出现 $p > 0.05$ 的结果），需要进一步区分是证据不足还是效应本身不存在。然而，常用的原假设显著性检验(Null hypothesis significance test, NHST)无法提供支持零效应的证据。近年来，等价检验、贝叶斯估计和贝叶斯因子三种方法逐渐被用于评估零效应。文章介绍了三种方法的原理，并通过两个实例分析，展示三种方法的实际应用。三种方法各有特点：等价检验在逻辑上是对 NHST 的拓展，易于从传统统计中延伸使用；贝叶斯因子的解读符合直觉，逻辑上清晰；贝叶斯估计则具有较强的灵活性，可拓展于更多的研究问题。三种评估零效应的方法，可能能够帮助心理学研究者在实际研究中进行合理的统计推断和研究决策。

关键词 零效应；p 值；等价检验；贝叶斯估计；贝叶斯因子

1 引言

原假设显著性检验（Null hypothesis significance test, NHST，也翻译为零假设显著性检验或者虚无假设显著性检验）是目前使用最为广泛的统计推断方法。在 NHST 框架下，研究者通常在假定原假设为真的前提下，根据 p 值是否小于预先设定的 α（如：α = .05）决定是否拒绝原假设(Wasserstein & Lazar, 2016)，进而做出是否接受备择假设的统计推断。然而，p 值大于 α 的结果（即不显著的结果）并不能作为支持原假设（Null hypothesis, H_0）的证据(Greenland et al., 2016; Wasserstein & Lazar, 2016)。正是由于 NHST 的理论前提是假定原假设为真，因此，无论 p 值是否小于 0.05，均无法评估原假设本身是否为真。也就是说，如果研究者的原假设是零效应(Null effect)，即“效应量为零”或者“效应不存在”时，无论 p 值大小如何，均无法评估零效应。

在实际研究中，研究者经常需要评估零效应。例如，在一些实验组/控制组匹配的研究设计中，研究者需要对无关变量进行匹配（如：两组被试的性别、年龄），即希望通过统计推断得到“两组被试在这些无关变量上没有差异”的结论。又如，某些理论的假设可能在特定情况下效应不存在，证实零效应可为这些理论提供支持。另外，当不显著结果与研究者的预期不符时，研究者同样需要合理评估支持零效应的证据强度，从不显著结果中获取更多有效信息，帮助研究决策(Harms & Lakens, 2018)。

由于能够有效评估零效应的统计方法在心理学研究中鲜有提及，许多研究者错误地使用不显著结果来支持零效应(Amrhein, Greenland, & McShane, 2019; Gigerenzer, 2004, 2018; Greenland et al., 2016; X. Lyu, Xu, Zhao, Zuo, & Hu, 2020; Z. Lyu, Peng, & Hu, 2018)。Lyu 等人(2020)的调查发现有超过半数（54%）的心理系学生或研究者将 $p > .05$ 解读为“证实了原假设”。对已发表论文的分析也表明，研究者易将“$p > .05$”的结果作为“支持零效应”的证据(Aczel et al., 2018; 王珺等, 2021)。对不显著结果的错误解读可能会直接导致统计推断出现偏差。例如，匹配组研究中，对年龄进行独立样本 t 检验后得到 $p > .05$，即使结果发现两组差异的效应量 Cohen’s d 很小（如小于 0.3），也并不能通过统计推断得到两组被试的年龄无差异（或等价）的结论，此时如果推断组间年龄没有差异则可能导致对实验操纵效应的错误推断。另外，忽视对 NHST 下不显著结果的进一步分析，错误地认为所有的不显著结果都没有发表价值，会进一步加深发表偏见(Forstmeier, Wagenmakers, & Parker, 2017)。

综上，研究者需要合适的统计方法来评估零效应。近年来，研究者提出了三种可以用来评估零效应的方法——等价检验(Equivalence test)(Meyners, 2012; Rogers, Howard, & Vessey, 1993)、贝叶斯估计(Bayesian estimation)(Kruschke, 2011; McElreath, 2020)和贝叶斯因子(Bayes factor)(Aczel et al., 2018; 胡传鹏, 孔祥祯, Wagenmakers, Ly, 彭凯平, 2018)。本文将介绍三者的原理，并结合两个实例来讨论并对比三者的特点。

2 等价检验、贝叶斯估计和贝叶斯因子的原理

评估零效应的思路主要有两种。一种思路是设定了一个足够小的，几乎可以认为效应为零的区间，用于评估零效应(Meyners, 2012; Rogers et al., 1993)。这一区间即为“最小感兴趣的效应量区间”，也简称为“最小感兴趣区”（Smallest effect size of interest, SESOI）。目标效应量在 SESOI 内时，研究者可以认为效应量几乎为零，可以忽略不计。采用这种思路进行统计推断的方法有两种，分别是频率统计框架下的等价检验和贝叶斯统计框架下的贝叶斯估计。另一种思路，如贝叶斯因子所采用的，则回避效应量是否为零的问题，比较假定效应量为零的原假设与假定效应量不为零的备择假设在当前数据下出现的可能性，从而推断当前数据更支持哪个假设。

2.1 等价检验

等价检验从传统 NHST 扩展而来，目的是评估当前效应量是否足够小。等价检验的逻辑来源于最小效应量检验（Minimal-effects test）(Murphy, Myors, & Wolach, 2014)。NHST 是将效应量与零做比较，判断当前数据在假定效应为零（H_0）的情况下出现的概率是否足够小，从而推断是否拒绝原假设（图 1A）。如果研究者将 H_0 设定为一个区间，比如[-0.1, 0.1]，拒绝原假设则要求效应量要么显著大于 0.1，要么显著小于 -0.1（图 1B），需要进行两次单侧检验。这种做法被称为最小效应量检验。

等价检验则正好将最小效应量检验的 H_0 与 H_1 所对应的效应区间对调，H_1 在区间之内，而 H_0 在区间之外(Lakens, McLatchie, Isager, Scheel, & Dienes, 2018; Lakens, Scheel, & Isager, 2018)。如果 SESOI 为[-0.1, 0.1]，等价检验的原假设是效应量要么大于 0.1，要么小于 -0.1 的区间（图 1C），即“存在有意义的效应”；其备择假设是效应量在[-0.1, 0.1]之间，即效应量太小而可以认为“不存在有意义的效应”。如果当前数据拒绝了原假设，则可以接受备择假设，即“不存在有意义的效应”。

等价检验中的原假设和备择假设除了与传统 NHST 的原假设和备择假设具有不同的意义之外，其对于原假设的设定要求更高。相对于 NHST 中原假设假定效应量为零，在等价检验中，研究者需要指明的是原假设的范围，即备择假设（SESOI）之外的区间。结合已有研究和实际情况，SESOI 的设定有特定的方式（详见补充材料：osf.io/6mzr9），且必须有合理的原因。

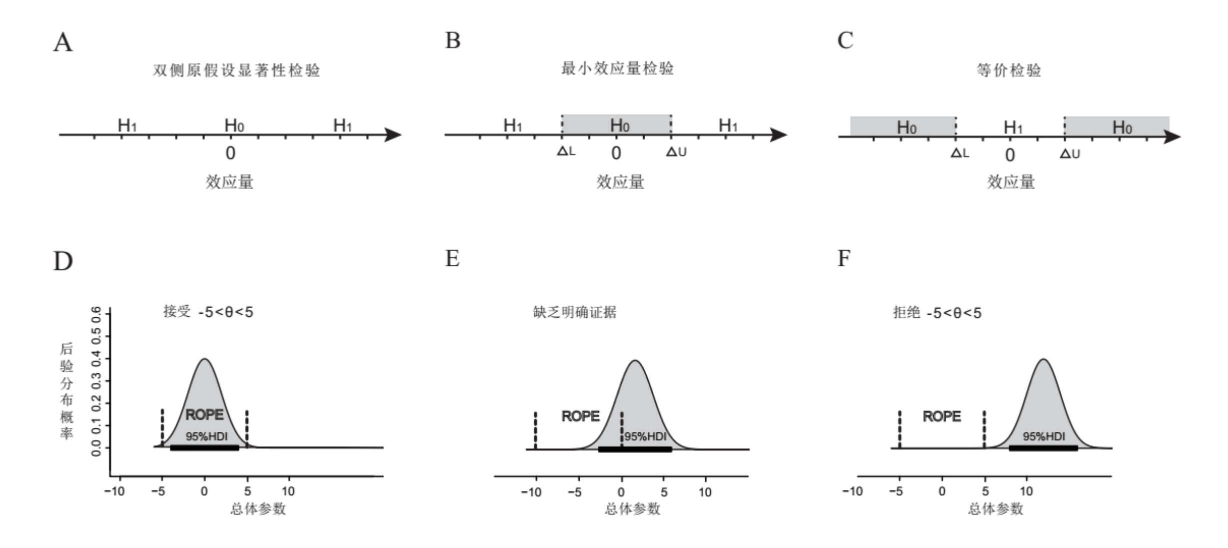


图 1. 等价检验和贝叶斯估计的原理示意图。（A）传统原假设显著性检验；（B）最小效应量检验；（C）等价检验。ΔL 表示最小感兴趣区（SESOI）的下限，ΔU 表示 SESOI 的上限；H_0：原假设，H_1：备择假设。贝叶斯估计推断中，结合最高密度区间（HDI）和实际等价区（ROPE）评估效应量 θ 的可信程度。评估有三种可能的结果，分别是接受零效应（D）、难以做出明确判断（E）和拒绝零效应（F）。

实际检验过程中，等价检验需要将实际数据与 SESOI 的下限 ΔL 和上限 ΔU 分别进行单侧的显著性检验，即两次单侧检验（Two one-side tests, TOST）。一次单侧检验的原假设是当前数据的效应量小于 SESOI 的下限 ΔL；另一次单侧检验的原假设则是当前数据的效应量大于 SESOI 的上限 ΔU。最后结合两个单侧检验的统计结果进行等价检验的推断：当且仅当 TOST 中的两个 p 值均小于 α 水平时，依据 NHST 框架的逻辑拒绝原假设，可以接受备择假设（“不存在有意义的效应”）。此时研究者可以认为存在统计上的等价性结果，即此效应足够小，在这一研究群体中是可以忽略的。但只要 TOST 中有一个 p 值大于 α 水平，就无法拒绝原假设（“存在有意义的效应”），即统计结果不能支持等价的结论(Lakens, Scheel, & Isager, 2018)。

值得注意的是，等价检验也可以通过基于参数估计的方法实现。频率统计框架下，研究者可以估计效应的值及其置信区间(王珺等, 2019)，然后根据效应量置信区间与 SESOI 重合的比例进行推断(Tryon, 2001)。

2.2 贝叶斯估计的原理

相比于基于频率学派统计的等价检验，贝叶斯估计是基于贝叶斯学派统计的一种评估零效应的方法。贝叶斯统计(Bayesian statistics)与频率统计(Frequentist statistics)的主要区别在于对概率(probability)的理解。频率中的概率表示在无数次的重复抽样中对于频率(frequency)的期望，即长期行为表现的结果。而贝叶斯统计中的概率表

示基于已有的信息，发生当前事件的可信程度(credibility)(Kruschke, 2014; McElreath, 2018)。具体到推断统计中，频率统计认为总体参数为固定值，而贝叶斯统计认为总体参数是对应概率分布下的随机取值，并且概率分布可以随着数据的获取而不断更新。贝叶斯统计的核心是贝叶斯法则(Bayes rules)。如果我们为了估计某一总体分布的参数(θ)而抽取了一定样本或“数据”(data)，基于贝叶斯法则可以得到下述公式：

$$P(\theta|data) = \frac{P(\theta)P(data|\theta)}{P(data)} \tag{1}$$

其中，$P(\theta|data)$表示基于数据得到的未知参数对应的概率分布，即后验分布(posterior distribution)；$P(\theta)$表示在获得数据前对于参数取值的信念，即先验分布(prior distribution)；$P(data|\theta)$表示当参数值为θ时，当前数据的概率或概率密度，即似然性(likelihood)；$P(data)$表示的是所有可能参数下出现当前数据的似然性的总合。在给定先验分布和数据的似然性之后，得到的后验分布表示同时考虑先验信息和数据表现的情况下总体参数的概率分布。简而言之，贝叶斯统计可以随着数据的累积不断更新后验，进而改变对参数不同取值的可信度(Kruschke & Liddell, 2018)。

应用贝叶斯估计评估零效应时，通过比较效应为零时的参数概率分布与后验分布下参数概率分布的差异，进行统计推断(Kirkwood & Westlake, 1981; Rouder, 2014; Westlake, 1976)。这里的后验分布下参数概率分布使用最高密度区间(highest density interval, HDI)表示，而效应为零时的参数概率分布是研究者预先设定的实际等价区(region of practical equivalence, ROPE)(Kruschke, 2014)。ROPE 类似于前文介绍的等价检验中 SESOI，是一个包括零的几乎可以忽略的效应区间。确定 ROPE 后，可以考察参数后验分布的 95%HDI 与 ROPE 的重合度来评估零效应。评估会出现三种不同的情况：接受零效应（图 1D）、拒绝零效应（图 1F）或者难以做出明确判断（图 1E）。具体而言，当 95%HDI 完全落在 ROPE 之内时，说明可能性最高的参数实际上等价于 0，因此可以接受零效应；当 95%HDI 和 ROPE 部分重合时，意味着只有部分可能性高的参数取值等价于 0，从而并不能做出明确判断；当 95%HDI 完全落在 ROPE 之外时，说明可能性最高的参数全部都不等价于 0，因此可以拒绝零效应(Kruschke, 2011)。总之，研究者可以将 HDI 与围绕零效应建立的 ROPE 进行比较以评估零效应。

值得注意的是，贝叶斯估计本身是基于数据进行模型拟合的过程，因此研究者可以使用不同的先验[1]和不同的模型。在这个过程中，需要考虑先验分布设定的合理性以及 MCMC 抽样收敛(convergence)，具体可以参考 Depaoli 和 Schoot (2017)。

[1] 值得注意的是，先验的设置是否合理有时难以判断，尤其是先验设定对后验的影响上。因此，对先验进行先验预测检验(Prior predictive check)也非常重要的，有兴趣的读者可以参考 McElreath (2020)的 *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2nd ed.)一书的第四章。

2.3 贝叶斯因子的原理

贝叶斯因子虽然也属于贝叶斯统计，但其在评估零效应时的思路与贝叶斯估计不同。其基本思路是通过模型比较的方式，获得给定数据下不同模型相对的可信程度。它尝试回答的问题是当前数据相对地更符合哪个模型。这里的模型对应于 NHST 中，即 H_0 模型或 H_1 模型。上文式（1）中的$P(data|\theta)$除了表示基于参数的先验分布得到当前数据的似然性，还可以理解成目标模型 H_0 或 H_1 为真的时候，出现当前数据的概率。而贝叶斯因子就是以这两者的比值定义的(Keysers, Gazzola, & Wagenmakers, 2020; Wagenmakers et al., 2018)：

$$\mathrm{BF}_{01} = \frac{P(data \mid H_0)}{P(data \mid H_1)} \tag{2}$$

其中，BF_{01}的下角标中 0 在前，1 在后，表示 BF_{01} 为 H_0 相对于 H_1 的贝叶斯因子。反之，BF_{10} 就是将式（2）中的分子分母颠倒，表示 H_1 相对于 H_0 的贝叶斯因子。当我们计算得到 $\mathrm{BF}_{01} = 9$ 时，表示当前数据出现在 H_0 为真的情况下的概率是出现在 H_1 为真的情况下的概率的 9 倍。得到贝叶斯因子之后，我们可以依据其大小得到数据支持两个模型的相对强度的证据。关于贝叶斯因子的解释，可以参考 Lee 和 Wagenmakers (2013) 基于 Jeffreys (1961)的解释提出的结果分类陈述。例如，BF_{01} 在[3, 10]之间时，可以解读为当前数据提供了中等强度的证据（Moderate evidence）来支持原假设（H_0）。

作为贝叶斯统计推断的一种方法，贝叶斯因子同样涉及先验的选择。一般根据先前研究确定先验，比如使用元分析得到的效应量及其对应的分布作为先验。而对于没有相关元分析的原创性研究，更常见的做法是使用一个标准化的先验，比如在贝叶斯 *t* 检验中，用柯西分布作为备择假设的先验(Rouder, Speckman, Sun, Morey, & Iverson, 2009)，$\delta \sim \mathrm{Cauchy}(x = 0, \gamma = 1)$。为了让备择假设的先验更符合现实，常用的计算贝叶斯因子的 R 包 BayesFactor 将默认的先验设定为 Cauchy(0, 0.707)。

4 等价检验、贝叶斯估计、贝叶斯因子的应用和比较

接下来，我们采用两个真实的数据来演示以上三种方法的应用。这两个例子在 NHST 框架下均采用独立样本 *t* 检验，且 *p* 值未达到显著水平。我们分别采用等价检验、贝叶斯估计和贝叶斯因子的方法重新对两个数据进行分析，并从评估零效应的能力、是否用到 SESOI/ROPE、是否报告不确定信息和可拓展性方面比较了三种方法。分析使用了 R 统计软件包 4.0.2 (R-Core-Team, 2019)。其中，等价检验使用的是 TOSTER 工具包(Lakens, 2017)，贝叶斯估计采用 BEST 工具包(Kruschke & Meredith, 2020)，贝叶斯因子采用 BayesFactor 工具包(Morey & Rouder, 2018)。两个实例的分析结果为典型的两种情况。其中，实例 1 展示的是数据无较强证据支持零效应的情

况，而实例 2 展示的是数据相对较强地支持零效应的情况。分析涉及的所有的数据、代码、结果及其解释见 osf.io/54qpv/。

4.1 实例 1：Kitchen Rolls

实例 1 的数据来自 JASP(jasp-stat.org)分析软件的示例数据“Kitchen Rolls”。该数据源自 Wagenmakers 等(2015)对 Topolinski 和 Sparenberg (2012)进行的重复研究。原研究的第二个实验中，两组被试分别以顺时针方向（$N_1 = 30$）和逆时针方向（$N_2 = 30$）拨动卷纸，然后填写一个测量开放性的问卷。结果发现，相比于逆时针拨动的被试，顺时针拨动的被试的开放性得分更高，$t(58) = 2.21, p < .031, d = 0.58$。Wagenmakers 等(2015)在预注册之后，重复了该研究的实验二。研究的数据包含两组被试在开放性人格特质上的得分，其中一组被试在填写问卷前顺时针旋转桌面上的纸卷（$N_1 = 48$），而另一组则在填写问卷前逆时针旋转纸卷（$N_2 = 54$）。我们采用 NHST、等价检验、贝叶斯估计和贝叶斯因子四种统计方法的双侧独立样本 t 检验来分析该数据，以评估零效应。由于等价检验和贝叶斯估计在统计过程中需要用到 SESOI 或 ROPE，因此首先确定 SSEOI。本分析参考 Simonsohn (2015)提出的重复研究中确定 SESOI 边界的方法，将 SESOI 的等价边界设置为原研究样本量之下，33%检验力可探测到的效应量，即 SESOI 为 [-0.40, 0.40]（计算过程见在线 R Notebook，osf.io/gn2hm/）。

NHST 未发现两组被试在开放性上的得分差异达到统计显著，$t(100) = -0.75$，$p = .453$，即未能拒绝原假设，但也无法提供支持零效应的证据。贝叶斯因子则为零效应提供了中等强度的证据，$BF_{01} \in (3, 10)$，具体而言，不同先验——Cauchy (0, 0.707)、Cauchy (0, 1)、Cauchy (0, 1.5)——之下的贝叶斯因子分别为 $BF_{01} = 3.71$、5.02、7.31。等价检验和贝叶斯估计的结果基本一致，即证据不足，无法判断数据是否支持零效应。具体表现为，在贝叶斯估计中，95%HDI 和 ROPE 部分重合，在等价检验中，TOST 左侧的 p 值大于 α 水平，因此均无法拒绝原假设（图 2A）。综合三种方法，可认为该数据无法为零效应提供较强的证据，同时也无法为效应的存在提供较强的证据。这表明，研究者需要进一步判断实验设计或者数据分析中可能存在的问题，并进行下一步研究和分析。

4.2 实例 2：Sociometric status and well-being

实例 2 的数据来自 Many Labs 2 项目（osf.io/uazdm/）中的一个研究。Many Labs 2 由 36 个国家和区域的不同实验室合力完成，共重复了 28 个经典的实验，总样本量达 15305(Klein et al., 2018)。实例 2 的数据来自报告中的第 12 个重复研究“Sociometric status and well-being”。该研究重复原研究中的实验三，探究相对于社会经济地位，

社会关系地位与幸福感的关系是否更紧密(Anderson, Kraus, Galinsky, & Keltner, 2012)。原研究中报告了一个显著的简单效应分析结果，相对低社会关系地位条件的被试，高社会关系地位条件的被试有更高的主观幸福感，$t(115) = 3.05$, $p = 0.003$, $d = 0.57$, 95% CI [0.20, 0.93]。Many Labs 2 主要重复了原研究中主观幸福感有差异的低社会关系地位条件和高社会关系地位条件，共采集了 6905 个样本。同实例 1，我们用四种统计方法下的双侧独立样本 t 检验分析了该数据。在分析之前，我们同样采用实例 1 的方式确定 SESOI 和 ROPE 为[-0.20, 0.20]。

NHST 未发现显著的效应，$t(6903) = -1.76$, $p = .08$，同样未能拒绝原假设，但也无法提供支持零效应的证据。然而等价检验、贝叶斯估计和贝叶斯因子的统计检验结果均支持了零效应（图 2B）。对于等价检验和贝叶斯估计，两组差异效应量的 90%CI 或 90%HDI 均完全落在 SESOI 和 ROPE 内。贝叶斯因子在 Cauchy (0, 0.707)、Cauchy (0, 1)、Cauchy (0, 1.5)三种先验分布下的结果分别为：$BF_{01} = 7.87$、11.11、16.64，达到了中等和较强程度支持零效应的证据(Lee & Wagenmakers, 2013)。其中，当先验分布的尺度参数变大时，BF_{01} 趋向于提供较强程度支持零效应的证据。三种评估零效应的方法一致支持了零效应，研究者可以较有信心地推断目标效应为零。

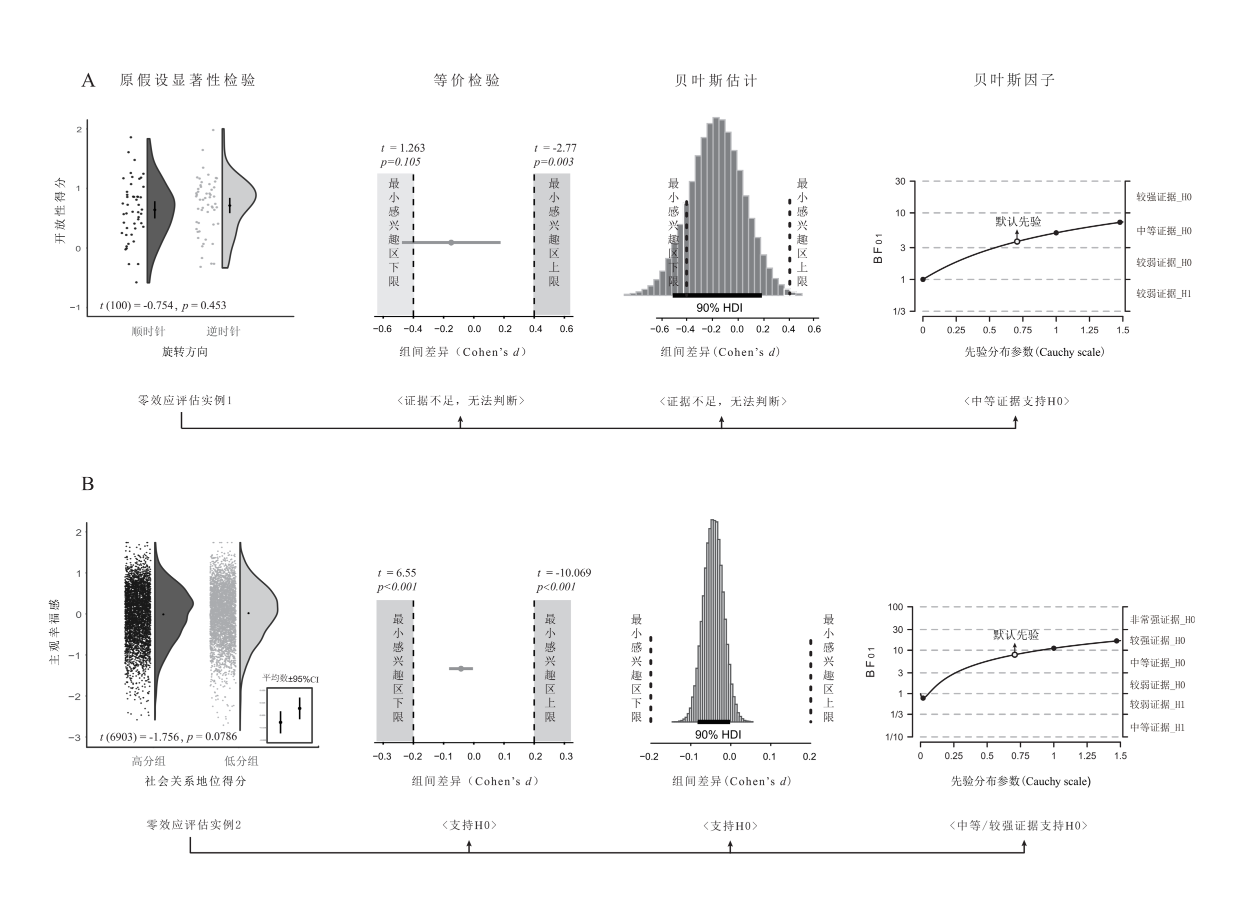


图 2．四种统计检验对两个实例数据的分析结果与推论。零效应评估实例 1（A）和实例 2（B）均用传统原假设检验、等价检验、贝叶斯估计和贝叶斯因子对数据进行分析，其中后三种方法均可以用来评估零效应。

4.3 等价检验、贝叶斯估计、贝叶斯因子的比较

在 NHST 框架下，以上两个实例数据均没有得到 $p < .05$ 的结果，即未能拒绝原假设。然而，这并不意味着当前数据可以证明零效应的存在。实例 1 的结果表明，虽然 NHST 得到的 p 值较大，但等价检验、贝叶斯估计、贝叶斯因子分析均表明该数据并不能为零效应提供较强的证据。而实例 2 的结果则表明，目标效应量与事先确定的近似于零的区间（SESOI/ROPE）无差别，而贝叶斯因子也提供了较强的支持零效应的证据，因此可以得到零效应的推论。两个实例数据的研究设计相对简单，因此三种方法均可以使用。但在更加复杂的研究设计中，是否能够同时使用三种方法可能需要进行深入地考察。以 TOSTER 包为例，等价检验目前只包括了 t 检验、元分析、相关分析等方法(Lakens, 2017)，这意味着其可拓展性方面存在限制。为了帮助研究者采用合适的方法，本文从几个维度对 NHST 和三种方法进行比较（表 1）。

表 1. 原假设检验、等价检验、贝叶斯估计和贝叶斯因子的特征及其对比。“O”表示有此特征，“X”表示无此特征。

特征	原假设检验	等价检验#	贝叶斯估计	贝叶斯因子*
能否拒绝零效应	O	X	O	O
能否支持零效应	X	O	O	O
是否用到最小感兴趣区（SESOI/ROPE）	X	O	O	X
是否报告不确定信息	X	O	O	X
可拓展性	高	低	高	中

此处对等价检验的可拓展性方面的评估主要基于当前可用的工具 TOSTER。

* 此处对贝叶斯因子的可拓展性评估主要基于 JASP 和 BayesFactor 工具包。

首先，等价检验、贝叶斯估计和贝叶斯因子均可以用来支持零效应，这是它们区别于 NHST 之处。因此，研究者在得到不显著结果时，可以采用这三种方法进一步从不显著结果中提取信息。其次，如果研究者希望支持零效应，使用等价检验与贝叶斯估计均需要使用 SESOI(Kruschke & Liddell, 2018; Lakens, Scheel, & Isager, 2018)，这意味着研究者需要提前确定一个合理的区间，才能进行合理的推断。但是计算贝叶斯因子时，则不需要确定 SESOI。第三，等价检验和贝叶斯估计提供了关于推断中不确定性的信息，且后者提供的不确定信息更为详实，描绘了参数的不同取值出现的相对概率(Kruschke & Liddell, 2018)；而贝叶斯因子未提供这些信息。第四，从可拓展性上来看，NHST 与贝叶斯估计均能够灵活地运用于各种统计推断的情境之中(Kruschke & Liddell, 2018;

Kruschke & Meredith, 2020)，但是贝叶斯因子和等价检验目前仍然较为限制。具体而言，贝叶斯因子目前主要可用于 t 检验、相关分析、方差分析和线性回归分析等常用的统计模型(Morey & Rouder, 2018)；而等价检验（基于 TOSTER）主要可用于 t 检验、元分析和相关分析(Lakens, 2017)。对于更加复杂的研究设计，如中介、调节分析等，贝叶斯因子和等价检验尚无可实现分析的代码。但是贝叶斯估计则能够应用于这些复杂的情境之中(Kruschke & Meredith, 2020)，如通过 R 工具包 brms 进行贝叶斯混合线性模型分析(Bürkner, 2017)。

除了三种方法原理特征上的差异外，随着样本量、等价区间的变化，三种方法的统计检验力（即效应量真值在等价区间内时，统计结果判断为等价的概率）也有不同。Linde, Tendeiro, Selker, Wagenmakers, 和 Ravenzwaaij (2020)通过一系列的模拟发现贝叶斯因子相对另外两种方法有更强的统计检验力，并且在样本相对较小的时候有更高的辨别力。

类似地，以上述两个实例的具体参数（样本量、等价边界）作为模拟参考，我们的模拟也发现，当效应量真值在区间 [0, 0.5]时，贝叶斯因子的统计检验力（即真实效应量落在等价区间，统计方法推断可以看作是效应量为零的比例）较高。但同样，其假阳性也更高（即真实效应量不在等价区间，但统计方法的结果认为其效应量可以看作为零的概率）（见图 3）。贝叶斯因子较高的敏感性在样本量小的时候更加明显，因此贝叶斯因子可能是小样本研究中用以支持零效应较好的方法，而适当收紧其判断标准（如将 $BF_{01}>10$ 作为等价标准，而非 $BF_{01}>3$）是权衡其较高统计检验力和较高一类错误的有效策略之一。

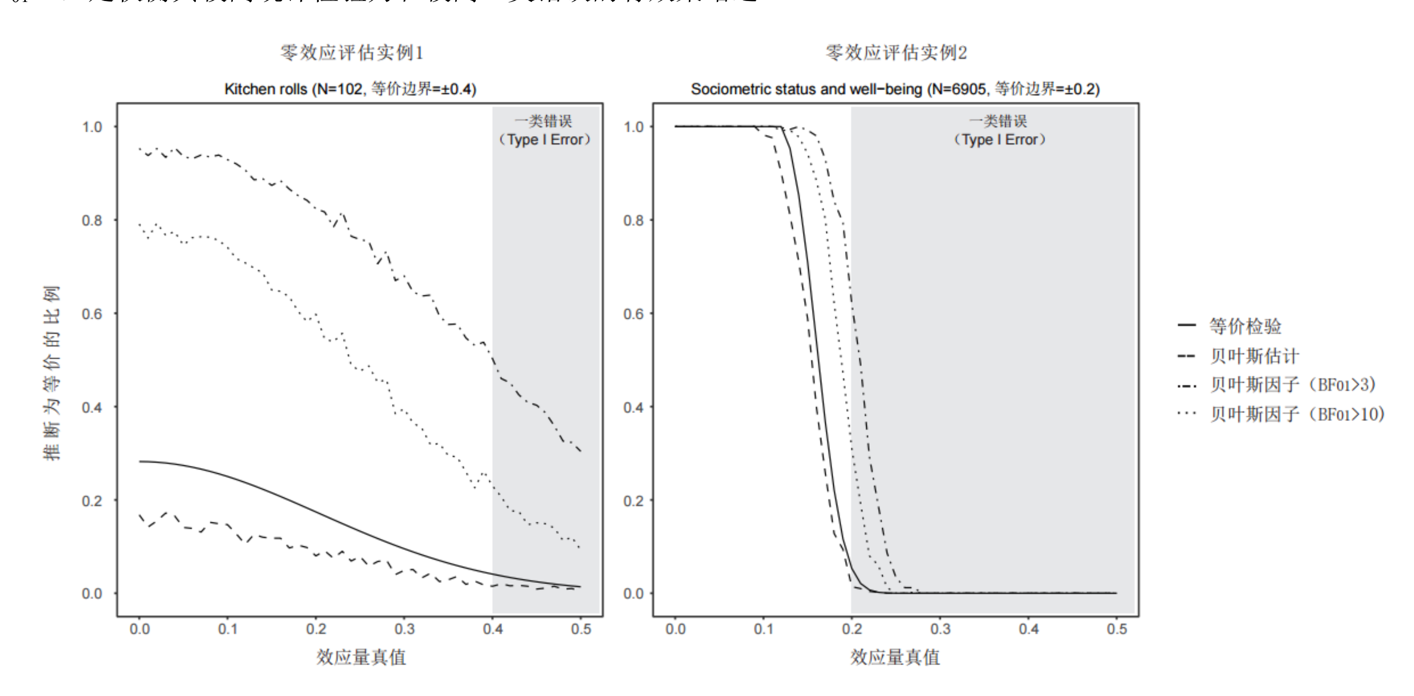


图 3. 等价检验、贝叶斯估计和贝叶斯因子在不同样本量、等价边界上的统计检验力及一类错误率。当效应量真值小于等价边界时，通过统计推断结果应判断为“等价”，即纵坐标推断为等价的比例反映的是三种方法的统计检验力；而当效应量真值大于等价边界时（图中阴影部分），纵坐标反映的则是三种方法的一类错误率。

三种方法相对于 NHST 均可以用于支持零效应，然而结果解释上存在理论上的区别。等价检验通过引入 SESOI 弥补了 NHST 功能上的缺陷，即不能用于推断效应不存在(Greenland et al., 2016; Wasserstein & Lazar, 2016)。其所在的统计框架仍为频率统计，即将统计推断建立在无数次的重复抽样中对于频率(frequency)的期望上。而基于贝叶斯统计框架下的贝叶斯因子和贝叶斯估计则有所区别。贝叶斯因子的统计推断本质上基于模型比较，即比较当前数据在两个相互竞争的模型中出现的相对概率(Keysers et al., 2020; Wagenmakers et al., 2018; 胡传鹏 et al., 2018)。贝叶斯估计则通过估计后验分布的 95%HDI 与类似于等价检验中 SESOI 概念的 ROPE 进行比较得到结论。推断的形式上，贝叶斯估计和等价检验相似，然而前者的 HDI 与后者的 CI 在对概率的认识上存在本质上的区别，也即贝叶斯统计和频率统计之间对概率不同认识上的区别(Kruschke, 2014; McElreath, 2020)。

5 总结与建议

心理学研究中不同的统计方法正在相互融合中共同发展(温忠粦, 方杰, 沈嘉琦, 谭倚天, 李定欣, 马益铭, 印刷中)，等价检验、贝叶斯估计和贝叶斯因子等统计方法的出现，一定程度上弥补了传统 NHST 无法评估零效应的缺陷，帮助研究者进一步区分“有证据支持零效应”和“没有证据支持有效应”这两种情况。三种方法在多个方面存在差异，各有特点，研究者可以根据当前研究的情况选择合适的方法。典型心理学研究情境下应如何进行选择合适的方法，可以参考补充材料中的流程图（补充材料，图 s1），但研究者需要在认真理解方法的基础上使用，避免滥用和误用(Gigerenzer, 2018)。

评估零效应的时候，以下三点值得注意：其一，如果采用等价检验和贝叶斯估计的方法，需要清楚地报告所采用的 SESOI/ROPE，并论证其合理性；如果采用贝叶斯估计或者贝叶斯因子，需要澄清所采用的先验及其合理性，也可以报告不同先验下的结果稳定性。其二，同时采用多种分析方法，交叉验证同一个结果可能是比较可靠的做法，例如上文的两个实例分别使用三种方法评估零效应。其三，我们建议在研究开始前或者数据分析前进行预注册。预注册中可以提供评估零效应的相应方法和参数，比如 SESOI/ROPE 和先验的确定。

参考文献

胡传鹏, 孔祥祯, Wagenmakers, E.-J., Ly, A., 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展, 26*(6), 951-965. doi:10.3724/SP.J.1042.2018.00951

王珺, 宋琼雅, 许岳培, 贾彬彬, 胡传鹏. (2019). 效应量置信区间的原理及其实现. *心理技术与应用, 7*(5), 284-296.

王珺, 宋琼雅, 许岳培, 贾彬彬, 陆春雷, 陈曦, . . . 胡传鹏. (2021). 解读不显著的结果：基于 500 个心理学实证研究的量化分析. *心理科学进展, 29*(3), 381-393. doi: 10.3724/SP.J.1042.2021.00381

温忠粦, 方杰, 沈嘉琦, 谭倚天, 李定欣, 马益铭. (印刷中). 新世纪 20 年国内心理统计方法研究回顾 [摘要]. *心理科学进展.* 取自 http://journal.psych.ac.cn/xlkxjz/CN/abstract/abstract5436.shtml

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., . . . Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science, 1*(3), 357-366. doi:10.1177/2515245918773742

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305-307. doi:10.1038/d41586-019-00857-9

Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect: Social status and subjective well-being. *Psychological Science, 23*(7), 764-771. doi:10.1177/0956797611434537

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1). doi:10.18637/jss.v080.i01

Depaoli, S., & Schoot, R. v. d. (2017). Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist. *Psychological Methods, 22*, 240-261. doi:10.1037/met0000065

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews of the Cambridge Philosophical Society, 92*(4), 1941-1968. doi:10.1111/brv.12315

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606. doi:10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198-218. doi:10.1177/2515245918771329

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. doi:10.1007/s10654-016-0149-3

Harms, C., & Lakens, D. (2018). Making'null effects' informative: statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research, 3*(Suppl 2), 382-393. doi:10.18053/jctres.03.2017S2.007

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *nature neuroscience, 23*(7), 788-799. doi:10.1038/s41593-020-0660-4

Kirkwood, T., & Westlake, W. J. (1981). Bioequivalence testing--a need to rethink. *Mathematics, 37*(3), 589-594. doi:10.2307/2530573

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443-490. doi:10.31234/osf.io/9654g

Kruschke, J. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299-312. doi:10.1177/1745691611406925

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Cambridge, Massachusetts: Academic Press.

Kruschke, J., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. doi:10.3758/s13423-016-1221-4

Kruschke, J., & Meredith, M. (2020). BEST: Bayesian estimation supersedes the t-Test. R package version 0.5.2. Retrieved from

https://CRAN.R-project.org/package=BEST

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*. doi:10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*: Cambridge university press.

Linde, M., Tendeiro, J. N., Selker, R., Wagenmakers, E.-J., & Ravenzwaaij, D. (2020). Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. doi:10.31234/osf.io/bh8vu

Lyu, X., Xu, Y., Zhao, X., Zuo, X., & Hu, C. (2020). Beyond psychology: prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology, 14*, e6. doi:10.1017/PRP.2019.28

Lyu, Z., Peng, K., & Hu, C. P. (2018). P-Value, Confidence Intervals, and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology, 9*, 868. doi:10.3389/fpsyg.2018.00868

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.): Chapman and Hall/CRC.

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference, 26*(2), 231-245. doi:10.1016/j.foodqual.2012.05.003

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-4.2. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*: Routledge.

R-Core-Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *psychological bulletin, 113*(3), 553-565. doi:10.1037/0033-2909.113.3.553

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301-308. doi:10.3758/s13423-014-0595-4

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. J. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237. doi:10.3758/PBR.16.2.225

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569. doi:10.1177/0956797614567341

Topolinski, S., & Sparenberg, P. (2012). Turning the hands of time: Clockwise movements increase preference for novelty. *Social Psychological and Personality Science, 3*(3), 308-314. doi:10.1177/1948550611419266

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*(4), 371.

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., . . . Sasiadek, A. (2015). Turning the hands of time again: a purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology, 6*, 494. doi:10.3389/fpsyg.2015.00494

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Boutin, B. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58-76. doi:10.3758/s13423-017-1323-7

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70*(2), 129-133. doi:10.1080/00031305.2016.1154108

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32*(4), 741-744. doi:10.2307/2529259

Evaluating null effect in psychological research: A practical primer

Abstract

In psychological researches, investigators need to provide not only evidence for the existence of effects, but also evidence for the non-existence of effects under some circumstances. However, the most widely used statistical inference framework in psychology, the null hypothesis significance test (NHST), cannot distinguish the evidence of absence from the absence of evidence. Here we introduced three methods, the equivalence test, Bayesian estimation, and Bayesian factor (BF), to Chinese researchers with two public datasets. Moreover, we compared these three methods from the following dimensions: whether a predetermined interval is needed; whether the test provided uncertainty information and whether the method is scalable in practice. By doing so, we provided practical tips for researchers who wish to apply these methods in their own researches. The current primer may help researchers to understand these methods for further application in their own research.

Keywords: Null effect; *p* value; Equivalence test; Bayesian estimation; Bayesian factor

补充材料

最小感兴趣区(SESOI)与实际等价区(ROPE)的确定

在等价检验和贝叶斯估计中，都会使用一个区间来定义一个足够小的，或者说可以被忽略的效应。在等价检验中，称为最小感兴趣区（SESOI），而贝叶斯估计将其定义为实际等价区（ROPE）。其他领域的研究者还会使用其他名称，如临床领域的临床等价区间(interval of clinical equivalence)(Lesaffre, 2008)和药理学的等价区间(equivalence interval)(Schuirmann, 1987)等。但这些概念本质上是相似的，都是为了定义一个包括零效应在内的足够小的区间，或者说更符合实际研究情况的零效应。由于 ROPE 与 SESOI 的相似，下文将仅从 SESOI 视角介绍。通过检验目标效应与该区间的相对关系可推断当前数据支持零效应、拒绝零效应还是无法做出判断(Kruschke & Meredith, 2020; Lakens, Scheel, & Isager, 2018)。当前数据的效应量区间一定时，如果 SESOI 比较宽松，则效应量区间可能完全落在 SESOI 内，得到支持零效应的推断；而 SESOI 范围较小时，效应量区间可能未完全在 SESOI 内，得到无法判断当前数据是否支持零效应的结论。因此 SESOI 的设定会直接影响零效应评估的结论。

SESOI 的设定需要具体问题具体分析。但是无论使用何种方法，研究者均需要对其设定合理性进行说明(Lakens et al., 2018)。通常，当研究者所感兴趣的效应量已经有先前研究进行过探索，则可以参考先前研究的结果。例如，Simonsohn (2015)建议，在重复研究中，可将 SESOI 的等价边界设置为之前研究的 33%检验力可探测到的效应。其理由在于，检验力低于 33%时得到的效应有多于 66%的概率得到的显著结果是不可信的(Simonsohn, Nelson, & Simmons, 2014)。但 Simonsohn (2015)的建议并非唯一的建议，Kordsmeyer 和 Penke (2017)则建议，在重复性研究中，可将 SESOI 的等价边界设定在先前研究的平均效应量上，并检验当前数据是否显著小于之前研究平均水平的效应量。然而这种方法无法排除先前研究随机性和出版偏见的影响。此外，还有观点认为可以将等价边界设定在之前研究正好可以观测到显著效应的临界值(Lakens et al., 2018)。另一个可能更稳健的方法是用元分析中估计效应量的置信区间（90%或 95%）的下边界（效应为正的情况下）作为等价边界(Perugini, Gallucci, & Costantini, 2014)。最后，值得注意的是，在频率学派和贝叶斯派两种不同的统计思想的框架下，SSEOI 和 POPE 对应的结果解释是有区别的(Kruschke & Liddell, 2018; Kruschke & Meredith, 2020)。

评估零效应的流程图

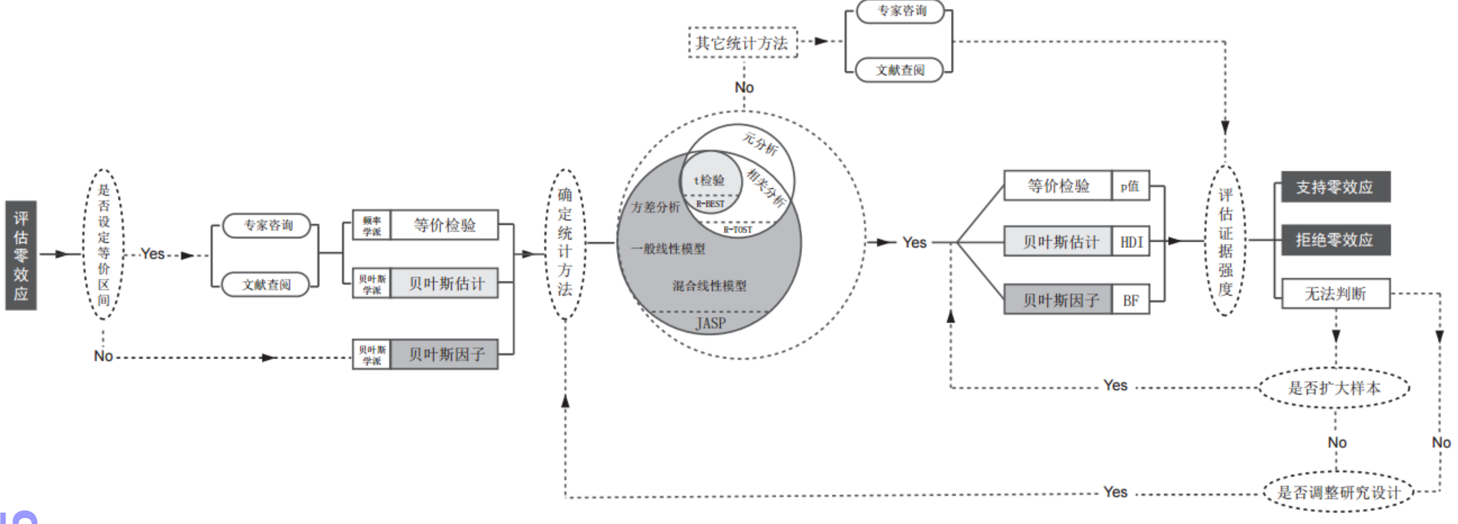


图 s1. 评估零效应的三种统计方法的使用流程。等价检验和贝叶斯估计在使用前需设定目标效应的等价区间，而贝叶斯因子不用；具体到对应的软件或编程语言，等价检验、贝叶斯估计和贝叶斯因子分别可以使用 R 中的 TOST 包、BEST 包和 BayesFactor 包实现，其中贝叶斯因子还可以使用 JASP 软件实现；三种方法分别根据各自特定的评估零效应的规则，得到支持零效应、拒绝零效应或无法判断的结论；若无法判断，研究者还可以考虑扩大样本量或调整实验设计，重新评估零效应。

补充材料参考文献

Kordsmeyer, T. L., & Penke, L. (2017). The association of three indicators of developmental instability with mating success in humans. *Evolution and Human Behavior, 38*(6), 704-713. doi:10.1016/j.evolhumbehav.2017.08.002

Kruschke, J., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. doi:10.3758/s13423-016-1221-4

Kruschke, J., & Meredith, M. (2020). Best: Bayesian estimation supersedes the t-test. R package version 0.5.2. Retrieved from https://CRAN.R-project.org/package=BEST

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Lesaffre, E. (2008). Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases, 66*(2), 150–154.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*(3), 319-332. doi:10.1177/1745691614528519

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657-680. doi:10.1007/bf01068419

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569. doi:10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. J. C. L. C. (2014). P-curve: A key to the file drawer. doi:10.1037/a0033242